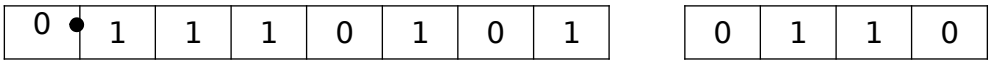
Worksheet 4 Floating point arithmetic **Answer**

Task 1

1. Convert the following floating point numbers from binary to decimal. Show your working.

(a)



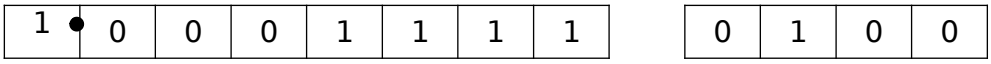
Mantissa Exponent

Mantissa: 0.1110101

Exponent = 6,

so move binary point 6 places right 0111010.1 = 58.5

(b)



Mantissa Exponent

Mantissa: 1.0001111

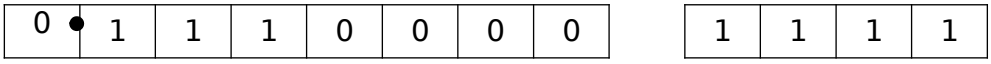
1s complement 0.1110000

Add 1 to get 2s complement 0.1110001

Exponent = 4,

so move binary point 4 places right 01110.001 = 14.125 Ansswer = -14.125

(c)



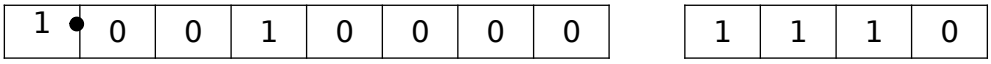
Mantissa: 0.1110000

Exponent = -1 (the negative exponent can be calculated using 2s complement or as -8 + 7, as sign bit has the value -8)

so move binary point left 1 place 0.0111000 = 0.0111000 =.25 + .125 +.0625

= .4375

(d)



Mantissa: 1.0010000

1s complement 0.1101111

Add 1 to get 2s complement (-) 0.1110000

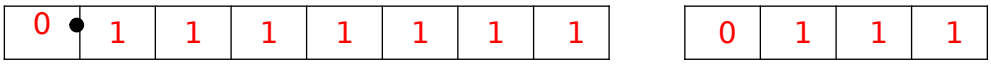
Exponent = -2,

PG ONLINE

so move binary point 2 places left (-) 0.0011100 = - (.125 + .0625 + .03125)

= - 0.21875

2. What is the largest number, in decimal that can be represented using this floating point system?



Exponent = 7 so move binary point 7 places right

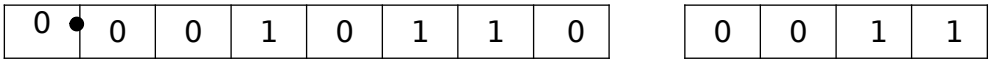
01111111 =127

Largest number = 127

Task 2

3. Convert the following binary numbers into normalised form:

(a)

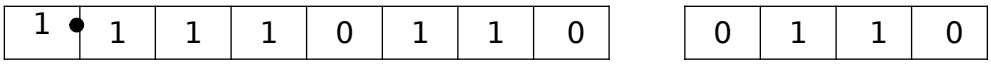


Move the binary point 2 places right so that there is a 1 directly after the sign bit

This makes the number larger, so subtract 2 from the exponent

The normalised number is 0.1011000 0001

(b)



Move the binary point three places to the right so that there is a zero directly after the sign bit

Subtract 3 from the exponent

The normalised number is 1.0110000 0011

4. Convert the following from decimal to normalised binary floating point, using an 8-bit mantissa and a four-bit exponent. show your working.

(a) 45.5

convert to binary 0101101.1

move point 6 places left 0.1011011 exponent 6

Answer 0.1011011 0110

(b) -14.5

convert 14.5 to binary 01110.100

one's complement 10001.011

two'complement 10001.100

move point 4 places left 1.0001100 exponent 4

Answer: 1.0001100 0100

5. What is the most negative number that can be held in an 8-bit mantissa and a 4-bit exponent? Express the answer as a normalised floating point binary number.

Answer:

1	0	0	0	0	0	0	0	0	1	1	1

Task 3

6. Add together the two normalised binary numbers shown below, giving the result in normalised floating point binary form.

0	1	1	0	0	1	0	0	0	0	1	0

0	1	1	0	0	1	1	0	0	1	0	0

Convert each number to fixed point binary:

11.00100

1100.110

Add: 1111.111

Move the binary point 4 places left, to give 0.1111111

Set exponent to 4.

the result is

0	1	1	1	1	1	1	1	0	1	0	0

7. Subtract the second binary number below from the first, giving the result in normalised floating point form.

0	1	1	0	0	1	0	0	0	0	1	0

0	1	0	1	1	0	0	0	0	0	1	1

Convert the numbers to fixed point form

first number (A) 011.0010 (=3.125)

second number (B) 0101.1000 (= 5.5)

one's complement of B: 1010.0111

\+ 1 1

two's complement (-B) 1010.1000

first number (A) 011.0010

-B + A 1101.1010 (two's comp = -2.375)

Normalise by moving binary point 2 places left 1.0110100 (makes number smaller)

Add 2 to the exponent

1	0	1	1	0	1	0	0	0	0	1	0

8. Subtract the second binary number below from the first, giving the result in normalised floating point form.

0	1	1	0	0	1	0	0	0	1	0	0

0	1	1	1	0	0	0	0	0	0	1	0

Convert the numbers to fixed point form

first number (A) 01100.100 (= 12.5)

second number (B) 00011.100 (= 3.5)

Find one's complement of the second number: 11100.011

\+ 1 1

two's complement (-B) 11100.100

first number (A) 01100.100

-B + A (1) 01001.000 (= 9.0)

Normalise by moving binary point 4 places left 0.1001000 (makes number smaller)

Set exponent to 4

0	1	0	0	1	0	0	0	0	1	0	0